图书情报工作
第65卷 第2期 2021年1月

外文期刊数据库内容质量指标的量化研究*

■ 庄纪林　陈凌　姚晓霞
北京大学图书馆　北京 100871

摘　要：[目的/意义]针对外文期刊数据库内容质量的双库与多库指标的量化，提出基于统一期刊空间的多库比较算法，实现任意多库之间的期刊识别和比对，提高这些指标的量化准确度。[方法/过程]在分析外文期刊数据库内容质量指标的内容及其复杂性的基础上，设计基于统一期刊空间的多库比较算法，并利用该算法计算出2018年DRAA中57个数据库的JCR刊数量和独有刊数量指标，作为算法有效性和准确性的验证。[结果/结论]作为各种电子资源评价指标系统的延伸研究，本研究能帮助图书馆优化自己的馆藏建设，使得外刊数据库的采购更加经济、合理。

关键词：外文期刊　数据库评价　内容质量　指标量化　统一期刊空间
分类号：G250.74
DOI：10.13266/j.issn.0252-3116.2021.02.006

1　引言

毋庸置疑，在移动阅读、云阅读遍地开花的时代，无论是读者对图书馆资源的利用上，还是图书馆在对资源的采购中，电子资源所占的比例越来越高，很多图书馆在电子资源的采购上所花费的金额已经超过了传统的纸质资源。特别在当前受新冠疫情的影响下，怎样利用有限的经费最大程度地保障电子资源的建设与服务，成为众多图书馆面临的迫切课题。

面对类型多样、价格不菲的电子资源，从内容质量、系统功能、使用情况、成本效益、数据库商服务、归档和长期保存等方面，很多研究人员提出了自己的电子资源评价指标，用来指导图书馆进行电子资源的建设。对于外文期刊数据库来说，内容质量指标评价的是数据库中期刊种类的数量和质量，如期刊数量、现刊数量/比例、JCR(Journal Citation Reports)/SNIP(Source Normalized Impact per Paper)期刊数量/比例、独有刊数量/比例等。因特网发展到现在，从硬件到软件，从系统到应用，电子资源服务平台的系统功能、性能、可靠性和易用性等方面都有了极大的保障，同时伴随着读者信息素养的普遍提高，读者不会对不同的数据库平台感到有明显的差异。特别是进入了大数据时代，数据为王的观念已经深入人心。内容质量成为了电子资源平台的核心竞争力，也是图书馆建设电子资源时的重中之重。

因此，对电子资源的内容质量进行准确的量化成为电子资源评价的首要任务。电子资源评价系统的指标，有些是定量的，有些是定性的，也有些是半定量半定性的。而内容质量指标一般都是定量的，这就为量化这些指标提供了可行性。但是，研究者一般都关注在应该建立哪些评价指标、采取何种方法筛选指标和设置权重、以及如何验证这些指标及权重是否合理等方面，却鲜有论文述及如何正确地量化这些指标。在实际操作时，大多依赖数据库提供商提供这些量化数据，但仅依靠数据库提供商提供的数据还是不够的，因为数据库提供商不会提供跨平台的数据库比较数据。针对这种情况，笔者基于高校图书馆数字资源采购联盟(Digital Resource Acquisition Alliance of Chinese Academic Libraries，DRAA)系统中的资源百科数据[1]，就外文期刊数据库内容质量方面的指标，提出一种通用的量化方法，作为各种电子资源评价指标系统的延伸研究，以便优化图书馆外文期刊数据库的采购，让有限

* 本文系国家社会科学基金项目“数字图书馆资源与服务绩效影响因素及评价体系研究”(项目编号：16BTQ031)研究成果之一。
作者简介：庄纪林(ORCID：0000-0002-7783-0964)，CALIS数据管理部副主任，副研究馆员，硕士，E-mail：zhuangjl@pku.edu.cn；陈凌(ORCID：0000-0002-2986-8976)，CALIS管理中心副主任，教授，硕士；姚晓霞(ORCID：0000-0002-3473-5377)，副馆长，研究馆员，博士。
收稿日期：2020-08-06　修回日期：2020-10-16　本文起止页码：54-64　本文责任编辑：徐健

的资金发挥最大的效用。

2 研究现状

随着图书馆越来越多地采购电子资源,电子资源在馆藏发展中占据着越来越重要的地位,电子资源的评估也越来越受到图书馆的重视。

国外的研究始于 20 世纪 90 年代后期,其评价对象通常设定为数字图书馆,以英美为主的国家和一些国际机构开展的项目为主。美国图书馆研究协会的"ARL"电子图书馆服务评价(E-METRICS)项目为 ARL 成员馆制定了一个网络统计和绩效测度数据收集手册[2];网络电子资源在线使用统计 COUNTER 项目规定了四大类数字资源的使用统计规范[3];欧洲图书馆绩效评价和质量管理系统 EQUINOX 项目也设计了自己的数字资源服务绩效评价指标体系[4];ISO2789 附录 A 设计了可获取性较强的图书馆电子服务使用评价指标[5];网络环境中国家公共图书馆统计和绩效测度项目发布了《公共图书馆网络化服务统计和绩效测度》手册[6];美国国家信息标准组织、美国国家图书馆和信息科学国家委员会和美国博物馆与图书馆服务协会共同修订了美国图书馆统计标准(ANSI/NISO Z39.7),通过了一系列图书馆统计标准和网络服务统计标准[7];以及各种基于 COUNTER 报告的采购决策实践研究[8-11]。

国内高校图书馆数字资源采购联盟为成员馆提供了基于 COUNTER 报告的使用统计;中国高等教育文献保障系统(China Academic Library & Information System,CALIS)论述了 CALIS 数字资源评估指标体系的评估对象和评估内容,列出了完整的指标体系及指标属性,并提出了应用指南[12];肖珑论述了电子资源评价指标体系的建立方法及其主要内容[13];姚晓霞全面介绍图书馆业绩评估国际标准 ISO11620,以及 EQUINOX 项目发展的电子图书馆业绩指标和 EQUINOX 系统[14];向英明探讨并建立了电子资源的综合评价指标体系及其数学模型[15];徐革运用专家问卷调查,进行了评价指标影响因子重要性程度和易获得性的调查研究[16];刘军运用专家打分法、层次分析法遴选评价指标并进行实证筛选[17];唐琼开展专家问卷调查,采用因子分析法提炼精简出关联度最强、最具代表性的电子资源选择标准[18];和媛媛针对基于用户满意度的电子资源评价指标体系,通过模糊多属性决策方法综合评价图书馆电子资源的利用情况[19];刘文梅以"211 工程"高校数字图书馆为研究对象,对已有研究提出的指标在我国的适用性进行了调查和分析[20];索传军参考相关的国际、国家标准及国内外研究项目成果,结合我国数据库使用统计数据的可获取情况,构建了数字馆藏服务绩效的评估指标体系[21];李新霞对国内外数字图书馆绩效评估在电子资源、计算机终端、经济成本和用户培训方面的网络服务统计指标体系和绩效评估指标体系进行了比较研究[22];张轶华在总结分析现有评价研究的基础上,结合上海交通大学图书馆的实践经验,将构建综合评价模型分解成评价指标遴选、数据采集及处理、指标权重设置 3 个重要环节[23]。

可以看出,以上研究大多注重于指标选择的指导性原则、指标的遴选、指标适用性的分析,但在指标值的获取方面关注较少,特别是在内容质量评价指标方面,由于指标数据获取的复杂性,往往只能依靠数据库商和电子资源分析平台提供的数据。依靠这种方式只能得到单库的指标,如果涉及到多库指标的计算,研究者所采用的方法不一而足,一旦比较对象改变,数据分析处理的方法就得重新设计,没有形成系统的指标量化方法。为了改变这种现状,笔者特别针对外文期刊数据库的内容质量这一指标下面的二级指标,提出了统一期刊空间中的多库比较算法,把待分析的数据库中的期刊按 ISSN、EISSN 和刊名分层映射到统一期刊空间中。而在进行刊名映射时,先通过刊名规范化提取刊名的特征进行待选期刊的圈定,然后再利用 N-Gram 算法根据原刊名在待选期刊中选出最佳匹配期刊。本研究中所应用的刊名规范方法是笔者在研究了大量不同库的刊名书写形式的基础上,经过长时间的积累总结提炼出来的,保证了刊名特征提取的精炼性和简洁性。而 N-Gram 算法的引入使得最佳刊名的快速匹配成为可能。

3 内容质量指标量化分析及算法设计

3.1 外文期刊数据库内容质量指标

国外文献在电子资源评价指标的遴选方面一般只提供指导性原则,实际工作中大都利用 COUNTER 报告指导图书馆进行电子资源的采购。国内的研究人员则在指标的完整性和系统性方面做了充分的扩展,也在一定程度上达成了共识。肖珑和张宇红在《电子资源评价指标体系的建立初探》一文中提出了电子资源评价体系的 6 个一级指标:电子资源内容、检索系统及功能、使用情况、价值与成本核算、出版商/数据库商的服务、存档[13]。之后相关研究大都按照这 6 个一级指标建立相应的指标体系。本研究着重于第 1 个一级指

标“电子资源内容”的量化分析方面。对于外文期刊数据库来说，指标“电子资源内容”对应于指标“内容质量”，评价的是数据库中期刊种类的数量和质量。其下重要的二级指标一般有：①期刊数量；②现刊数量/比例；③JCR/SNIP 刊数量/比例；④独有刊数量/比例。

分析以上指标，我们不难发现：

首先，这些指标的量化涉及到多种刊表。我们把用以评价的数据库称为源数据库，源数据库的刊表数据可从 DRAA 网站下载、从数据库服务平台下载或向数据库提供商索取。评价时用作参考的数据库，如 JCR 期刊库、SNIP 期刊库，称之为参考数据库，这些参考数据库也可分别从各自的网站下载；

其次，按量化时使用到的期刊数据库数量的不同，这 4 个指标又可分为单库指标、双库指标以及多库指标。其中，期刊数量、现刊数量/比例这两个指标的量化只涉及源数据库自身，属于单库指标；指标 JCR/SNIP 刊数量/比例的量化需要在源数据库与参考库之间进行期刊比对，属于双库指标；指标独有刊数量/比例的量化涉及到多个源数据库之间的期刊比对，属于多库指标。

除第 1 个二级指标之外，其余 3 个二级指标只要先量化出数量指标，再将此数量指标除以第一个指标“期刊数量”就可得到相应的比例指标值。

3.2 外文期刊数据库内容质量指标量化方法

单库指标的量化相对比较简单，直接在该源数据库的刊表中进行统计即可。以从 DRAA 网站得到的资源百科数据为例，其刊表至少包含以下字段：数据库名、期刊名、出版者、ISSN、EISSN、学科分类、主题词、收录范围、URL、PLATFORM、DOI、全文收录滞后期、全文收录起始日期、全文收录结束日期。利用这些字段可实现期刊数量、现刊数量/比例指标的量化，另外还可以得到一些其他比较重要的指标，如全文收录滞后期。当然，这些指标值的获取都是建立在数据准确、完备的情况下，如果相应的字段值有错误或者缺失较多，就会对指标计算造成一定的干扰。

从 2018 年 DRAA 集团采购的数据库中（以下简称 DRAA 2018）选取 57 个外文期刊源数据库，其中包括 43 个电子期刊库和 14 个全文数据库，统计其关于内容质量方面的一些单库指标，得到其相应的 Top 10 数据。同时，为了清楚地揭示电子期刊库和全文数据库各自的数据，进行了分组统计，列出电子期刊库的前 7 名和全文数据库的前 3 名，如表 1 – 表 6 所示：

表 1 期刊数量排名前 10 的外文期刊数据库统计

数据库类别	序号	数据库名称	期刊数量/种
电子期刊库	1	Journal Storage	3 986
	2	Elsevier ScienceDirect	2 369
	3	Taylor & Francis 期刊数据库	2 356
	4	Springer 电子期刊	1 953
	5	Wiley Online Library	1 446
	6	Project MUSE E-journal	639
	7	SAGE 现刊数据库	587
全文数据库	1	H. W. Wilson 数据库	20 322
	2	Academic Search Premier	16 387
	3	Academic Search Complete	15 810

表 2 期刊数量排名倒数 10 名的外文期刊数据库统计

数据库类别	序号	数据库名称	期刊数量/种
电子期刊库	1	PNAS 数据库	1
	2	Science Online	7
	3	SPIE Digital Library	10
	4	American Society for Microbiology	12
	5	The JAMA Network	12
	6	IWA	12
	7	SAE 国际汽车工程师学会数据库	16
全文数据库	1	World Bank	3
	2	EBSCO 食品科学全文数据库	867
	3	TWS 台湾学术期刊在线数据库	1 738

表 3 现刊数量排名前 10 的外文期刊数据库统计

数据库类别	序号	数据库名称	现刊数量/种
电子期刊库	1	Taylor & Francis 期刊数据库	2 356
	2	Springer 电子期刊	1 953
	3	Elsevier ScienceDirect	1 949
	4	Wiley Online Library	1 446
	5	SAGE 现刊数据库	587
	6	Project MUSE E-journal	562
	7	Oxford Journals Collection	380
全文数据库	1	H. W. Wilson 数据库	17 096
	2	Academic Search Premier	14 965
	3	Academic Search Complete	13 775

表 4 现刊数量排名倒数 10 名的外文期刊数据库统计

数据库类别	序号	数据库名称	现刊数量/种
电子期刊库	1	PNAS 数据库	1
	2	Science Online	6
	3	IMechE Journals Collection	7
	4	AIAA 美国航空航天学会数据库	8
	5	SPIE Digital Library	10
	6	ASME 数据库	12
	7	IWA	12
全文数据库	1	World Bank	2
	2	EBSCO 食品科学全文数据库	657
	3	TWS 台湾学术期刊在线数据库	854

庄纪林，陈凌，姚晓霞.外文期刊数据库内容质量指标的量化研究[J].图书情报工作,2021,65(2):54-64.

表5 现刊比例排名前10的外文期刊数据库统计

数据库类别	序号	数据库名称	现刊数量/种	期刊数量/种	现刊比例/%
电子期刊库	1	Taylor & Francis 期刊数据库	2 356	2 356	100.00
	2	Springer 电子期刊	1 953	1 953	100.00
	3	Wiley Online Library	1 446	1 446	100.00
	4	SAGE 现刊数据库	587	587	100.00
	5	Oxford Journals Collection	380	380	100.00
	6	Cambridge Journals Online	367	367	100.00
	7	Emerald 全文期刊库	266	266	100.00
全文数据库	1	Academic Search Premier	14 965	16 387	91.32
	2	EBSCO 环境科学全文数据库	1 772	1 953	90.73
	3	Academic Search Complete	13 775	15 810	87.13

表6 现刊比例排名倒数10名的外文期刊数据库统计

数据库类别	序号	数据库名称	现刊数量/种	期刊数量/种	现刊比例/%
电子期刊库	1	Journal Storage	0	3 986	0.00
	2	IMechE Journals Collection	7	25	28.00
	3	ASME 数据库	12	30	40.00
	4	AIAA 美国航空航天学会数据库	8	17	47.06
	5	IEEE/IET Electronic Library	256	503	50.89
	6	APS 全文电子期刊数据库	14	22	63.64
	7	BioOne 数据库	147	196	75.00
全文数据库	1	TWS 台湾学术期刊在线数据库	854	1 738	49.14
	2	ABI/INFORM Collection	4 229	7 099	59.57
	3	World Bank	2	3	66.67

双库指标和多库指标的量化比单库指标的量化要复杂一些。计算双库指标时，以计算某源数据库的JCR刊数量为例，假设该源数据库的期刊所组成的集合为A，JCR参考库中期刊所组成的集合为B，对A和B求交集得到集合C，然后求C的模就得到该源数据库的JCR期刊数量。在计算多库指标时，以计算源数据库的独有刊数量为例，假设该源数据库的期刊所组成的集合为A，其余源数据库中的期刊并集为B，差集A-B中的期刊就是独有刊，其模就是独有刊的数量。

在量化双库指标和多库指标时，最主要的困难在于期刊识别。相当一部分外文期刊数据库商提供的库刊表普遍存在ISSN和EISSN缺失、刊名和出版商等字段著录随意性较大的问题，从而导致期刊识别变得困难重重。在只有少数几个库参与比较时，解决问题的难度还不是很大，当几十上百个库需要进行不同组合的比较时，解决问题的难度就急剧上升。在这种情形下，如果没有一个统一的指导方法，双库指标和多库指标的量化几乎是不可能的事情。

3.3 多库指标量化的复杂性分析

（1）相当一部分外文期刊数据库商提供的库刊表存在ISSN和EISSN都缺失的情况。

以上述DRAA 2018数据为例，经统计共有57个源数据库，其中20个源数据库出现了ISSN和EISSN都为空值的情况，按照空值数由大到小的顺序排列，选取前5个源数据库，如表7所示：

表7 DRAA 2018中ISSN和EISSN字段均空的期刊记录

数据库名称	ISSN 和 EISSN 字段均空的期刊数/期刊总数/种
ABI/INFORM Collection	3 796/7 099
H. W. Wilson 数据库	2 578/20 322
Business Source Premier	573/6 208
Business Search Complete	573/6 208
TWS 台湾学术期刊在线数据库	366/1 738

即使在以高数据质量著称的JCR数据库中，以2018年的JCR刊表(2018年WoS中SCIE、SSCI与AHCI的合集，以下简称JCR 2018)为例，经统计，14 436条期刊数据中(SCIE合集9 238条，SSCI合集3 421条，AHCI合集1 777条)共有9条存在ISSN和EISSN都为空的情况。造成这种情况的可能性有多种，可能是本来就没有ISSN和EISSN，也有可能是著录人员的疏忽造成的。

（2）外文期刊数据库商提供的库刊表中刊名字段

缺少统一的规范，著录时的随意性较大。

以 DRAA 2018 中 Academic Search Premier 与 ProQuest Research Library 这两个数据库为例，选取这两个库中 ISSN 相同的期刊，比较这些 ISSN 相同的期刊是否具有相同的刊名，结果发现两库共有 2 610 条数据具有相同的 ISSN，其中 904 条数据具有不同的刊名，具体数据见表 8。

在表 8 中 904 条数据中选取了 12 条来展现刊名不一致的复杂情况，见表 9。

表 8　刊名著录随意性统计

数据库名称	期刊总数/种	ISSN 非空刊数/种	ISSN 相同刊数/种	ISSN 相同、刊名不同刊数/种
Academic Search Premier	16 387	16 607	2 610	904
ProQuest Research Library	6 111	5 601		

表 9　刊名著录随意性示例

ISSN	Academic Search Premier 库中刊名	ProQuest Research Library 库中刊名
8756 - 3894	*TechTrends: Linking Research & Practice to Improve Learning*	*TechTrends*
8755 - 4550	*Women & Language*	*Women and Language*
8755 - 4178	*Journal of Feminist Studies in Religion (Indiana University Press)*	*Journal of Feminist Studies in Religion*
8750 - 6874	*Informationweek*	*InformationWeek*
2470 - 9506	*South: A Scholarly Journal*	*South*
2411 - 7862	*Literacy Today* (2411 - 7862)	*Literacy Today*
2374 - 0663	*TD: Talent Development*	*Talent Development*
2234 - 8867	*Journal of East Asian Economic Integration (JEAI)*	*East Asian Economic Review*
2178 - 5198	*Acta Scientiarium: Education*	*Acta Scientiarum. Education*
2160 - 603X	*Campaigns & Elections* (2010)	*Campaigns & Elections*
1862 - 6254	*Physica Status Solidi - Rapid Research Letters*	*Physica Status Solidi. Rapid Research Letters*
1742 - 464X	*FEBS Journal*	*FEBS Journal, The*

表 9 只是列举了一部分刊名不一致的情况，实际的情形要复杂得多。对于具备 ISSN 或者 EISSN 的记录，只需通过 ISSN 或者 EISSN 进行期刊识别。对于 ISSN 和 EISSN 都缺失的记录，需要通过其他手段进行期刊识别。笔者采用刊名字段作为 ISSN 和 EISSN 缺失情况下的辅助手段。首先根据各种复杂的刊名不一致的情况，提炼出刊名规范化的规则。对刊名应用此规范化规则，相当于提取刊名特征。然后通过刊名特征进行相等匹配，刊名特征相同则是同一种刊。如果通过刊名特征匹配出现一对多的情况，则引进 N-Gram 算法，计算两个原始刊名之间的相似度，选择相似度第一高者为最佳匹配。

3.4　统一期刊空间中的多库比较算法

为了进一步减小多库指标量化的复杂度，笔者引入了统一期刊空间中的多库比较算法。该算法的主要特点是把源数据库和参考库中的期刊记录统一映射到某个期刊空间中，使得每种期刊都具有一个唯一的身份识别码，从而实现了任意多库之间的期刊识别和比对。

在计算双库指标和多库指标时，涉及到源数据库之间以及源数据库与参考库之间的期刊比对。以 DRAA 2018 为例，其中有 57 个源数据库，如果为这 57 个源数据库的双库和多库指标的量化都单独设计一套期刊识别和比对算法，其工作量将是巨大而繁琐的。

为此，这里提出了统一期刊空间中的多库比较算法。如果能够把各种库中的期刊统一映射到某个库中的 ID，然后基于这个 ID 进行期刊集合之间的运算，那么问题就简单多了。显然，这个用来进行期刊映射的数据库必须能最大限度地包含所有其他数据库的期刊，乌利希期刊数据库[24] 可以满足这个要求。乌利希期刊数据库收录了 15 万个出版商、39 万多种期刊的详细数据，堪称期刊数据大全，是最适合作为期刊比对的统一期刊空间的数据库。本算法将源数据库中的期刊记录按 ISSN、EISSN 和刊名字段（以及规范化后的刊名字段），映射到乌利希期刊数据库空间中，返回一个乌利希数据库中期刊 ID（以下简称 jid）给相应的期刊记录，然后各库之间只需使用该 jid 进行相应的集合运算，即可量化所需要的指标。

需要说明的是，乌利希数据库中的 jid 在乌利希网站的列表页面和详细页面中都是不可见的，当对选中的记录进行下载时，能在 TitleId 字段看到该 jid，如图 1 所示：

A	B	C	D	E	F	G	H	I	J
TitleId	标题	ISSN	媒体类型	状态	SubjectCo	PublicationI	频率	卷信息	最新卷
12522	Moving Image	0276-3494	纸本	已停刊	MOTION	1972-19??			

图 1　乌利希数据库中期刊下载文件中的 TitleId 字段(jid)

另外，在乌利希期刊数据库中，少数 ISSN 或 EISSN 在乌利希数据库中对应多条期刊记录，此时还需进行借助刊名进行相似性比较，取刊名相似度最高的 jid。如果这多个记录的刊名相似度一样，则取第一条记录的 jid。这样保证了不同源数据库中的同一 ISSN 或 EISSN 能映射到乌利希期刊空间中的同一条记录，尽管在乌利希期刊空间中对应此 ISSN 或 EISSN 有多条记录。

在进行映射时，如果 ISSN 或 EISSN 非空，则使用 ISSN 或 EISSN 的相等匹配在乌利希期刊空间中进行 jid 的映射。如果这两者都为空，则使用刊名进行匹配。据此，可以设计出将源数据库中的期刊记录映射到乌利希期刊空间的 findUlrichJID 算法，算法具体如下：

```
算法 1 findUlrichJID
findUlrichJID(journaltitle, issn, eissn):
  jid = null
  if issn! = null then
    jid = matchByIssn(journaltitle, issn)
  else if eissn! = null then
    jid = matchByEIssn(journaltitle, eissn)
  else
    jid = matchByTitle(journaltitle)
  return jid
```

使用刊名匹配时，算法 matchByTitle 采用的匹配策略并不是简单的刊名相等匹配。对于那些 ISSN 和 EISSN 字段都为空值的记录，算法 matchByTitle 利用规范化后的刊名进行相等匹配。为了保证刊名规范化规则能尽可能地适应各种同一期刊在不同数据库中刊名不一致的情形，笔者提炼出如下刊名规范化规则：

(1)删除刊名中小/中/大括号及其内部包含的内容；

(2)删除冒号、逗号、两边带空格的连字符及其后面的内容；

(3)把“&”替换成“ AND ”，即两边带空格的 AND；

(4)替换除数字、字母和不带空格的连字符之外的其他字符为空格；

(5)大写刊名中所有字母；

(6)合并刊名中连续多个空格为 1 个空格；

(7)删除刊名首尾处的空格。

经过以上 7 条规范化规则的处理，使得不同数据库中同一期刊的不同形式刊名保留相同的基本特征，也使得通过规范化的刊名直接进行相等匹配成为可能。

利用以上规范化的刊名进行映射后，如果在乌利希期刊空间中没有匹配到，则 jid 返回空；如果匹配到乌利希期刊空间中的 1 个 jid，则返回该 jid；如果匹配到多个 jid，则返回原刊名相似度最高的记录的 jid。matchByTitle 的形式化算法如算法 2 所示：

```
算法 2 matchByTitle
matchByTitle(journaltitle):
  ntitle = normalizeTitle(journaltitle) //获取源数据库期刊刊名的规范化刊名
  journalList = getJidListFromUrlich(ntitle) //从乌利希数据库中获取规范刊名相同的期刊列表
  if journalList.size() = = 0 then
    return null
  else //这里合并了匹配到 1 个和多个记录的情形
    minScore = 整数最大值 //得分越小，相似度越高，该变量用以保存最小得分
    minJid = －1 //用以保存相似度最高的期刊 ID
    foreach journal in journalList do
      score = ngram(journaltitle, journal.title)
      if score < minScore then
        minScore = score
        minJid = journal.jid
    return minJid
```

规范化刊名用来在乌利希期刊空间中快速圈定疑似期刊，然后通过 N-Gram 算法计算原始刊名相似度，选出最合适的 jid。N-Gram 算法又称 N 元模型，它的作用是用来评估两个字符串之间的相似度，这是一种模糊匹配手段。笔者设计的算法 N-Gram 返回一个 score 值，score 越小则两个刊名的相似度越高。其形式化算法如算法 3 所示：

```
算法 3 N-Gram
ngram(journaltitle1, journaltitle2, n): //n 为 N-Gram 中的 N，一般设为 2 或 3
  ss = " " //ss 为长度为 n－1 的以空格填充的字符串
  str1 = ss + upcase(journaltitle1) + ss //在大写化后的 journaltitle1 两边补充字符串 ss
```

```
str2 = ss + upcase(journaltitle2) + ss //在大写化后的 journaltitle2 两边补充字符串 ss
len1 = len(str1)
len2 = len(str2)
for i = 0…(len1 - n) do //从左至右,把 str1 按 n 个字符进行切分,并加入到集合中
    set1.add(str1[i:i + n])
for i = 0…(len2 - n) do //从左至右,把 str2 按 n 个字符进行切分,并加入到集合中
    set2.add(str2[i:i + n])
set3 = set1 & set2 //集合 set1 与集合 set2 求交集得集合 set3
n1 = len(set1) //n1 为集合 set1 的元素个数
n2 = len(set2) //n2 为集合 set2 的元素个数
n3 = len(set3) //n3 为集合 set3 的元素个数
score = n1 + n2 - 2 * n3
return score
```

在本研究中,将算法 N-Gram 中的参数 n 设为 3。

所有在乌利希期刊空间中没有映射上的期刊,包括通过 ISSN 和 EISSN 没有映射上的和通过刊名没有映射上的,都排除在后续的期刊比对之外。

4 双库及多库指标的量化

现以 DRAA 2018 中 57 个源数据库和参考库 JCR 2018(2018 年 WOS 中 SCIE、SSCI 与 AHCI 的合集)为例,展示其在乌利希期刊空间中的映射结果,并计算 DRAA 2018 中 57 个源数据库的 JCR 期刊数量/比例与独有刊数量/比例这两个关键指标。当然,计算的前提是先完成 DRAA 2018 库刊表与 JCR 期刊表到乌利希期刊空间的映射。

4.1 DRAA 2018 在乌利希期刊空间的映射

使用本研究中的统一期刊空间中的期刊映射算法,以乌利希期刊库作为统一期刊空间,对 DRAA 2018 进行 jid 映射,映射结果如表 10 所示:

表 10 DRAA 2018 在乌利希期刊空间中的映射结果

匹配结果	匹配条件	期刊数量/种
0 匹配	ISSN(有 ISSN)	5 447
	EISSN(无 ISSN,有 EISSN)	230
	规范化刊名(无 ISSN,无 EISSN)	4 574
唯一匹配	ISSN(有 ISSN)	88 844
	EISSN(无 ISSN,有 EISSN)	3 519
	规范化刊名(无 ISSN,无 EISSN)	822
多个匹配,取最佳刊名匹配	ISSN(有 ISSN)	2 353
	EISSN(无 ISSN,有 EISSN)	25
	规范化刊名(无 ISSN,无 EISSN)	3 766
期刊总计		109 580

经统计,DRAA 2018 到乌利希期刊空间的映射率约为 91%,大约有 9% 的记录没有映射上,如图 2 所示:

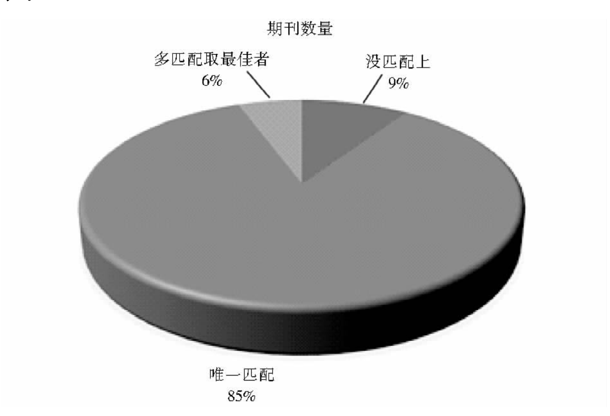


图 2 DRAA 2018 在乌利希期刊空间的映射率

没有映射上的最大原因可能有 3 个:①DRAA 2018 中存在错误记录;②乌利希期刊数据库存在错误记录;③乌利希期刊数据库中的期刊记录不全。

4.2 JCR 2018 在乌利希期刊空间的映射

运用同样的算法,得到了参考库 JCR 2018 在乌利希期刊空间中的映射结果,如表 11 所示:

表 11 JCR 2018 在乌利希期刊空间中的映射结果

匹配结果	匹配条件	期刊数量/种
0 匹配	ISSN(有 ISSN)	34
	EISSN(无 ISSN,有 EISSN)	4
	规范化刊名(无 ISSN,无 EISSN)	2
唯一匹配	ISSN(有 ISSN)	13 939
	EISSN(无 ISSN,有 EISSN)	328
	规范化刊名(无 ISSN,无 EISSN)	2
多个匹配,取最佳刊名匹配	ISSN(有 ISSN)	120
	EISSN(无 ISSN,有 EISSN)	2
	规范化刊名(无 ISSN,无 EISSN)	5
期刊总计		14 436

经统计,JCR 2018 在乌利希期刊空间中的匹配率达到了约 99%(见图 3)。这么高的匹配率也说明了 JCR 期刊数据正确率很高,同时也说明了 JCR 期刊很重要,基本都被乌利希期刊数据库收录。

4.3 DRAA 2018 中各库的 JCR 期刊数量和比例

现在,可以使用以上的映射结果,量化 DRAA 2018 中各源数据库的 JCR 期刊数量和比例。其主要算法就是先求出 DRAA 2018 中各库的 jid 集合与 JCR 2018 的 jid 集合的交集,然后对交集求模,得到 DRAA 2018 中各源数据库的 JCR 期刊数量,再把 JCR 期刊数量除以源数据库期刊总数,即得该源数据库的 JCR 期刊比例。

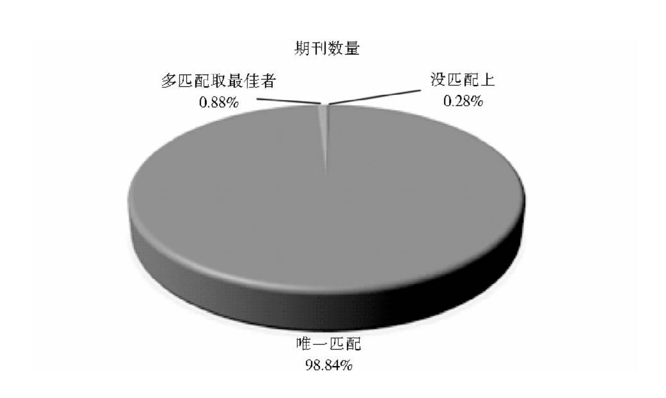


图 3　JCR 2018 在乌利希期刊空间的映射率

限于篇幅,表 12 和表 13 分别列出了 JCR 期刊数量正数前 10 名和倒数前 10 名的源数据库,表 14 和表 15 分别列出了 JCR 期刊比例正数前 10 名和倒数前 10 名源数据库。同样,为了清楚地揭示电子期刊库和全文数据库各自的数据,进行了分组统计,并按 7:3 的比例分别展示电子期刊库与全文数据库的结果。

表 12　DRAA 2018 中 JCR 期刊数量正数前 10 名的源数据库统计

数据库类别	数据库名	JCR 期刊数量/种	期刊总数/种	JCR 期刊比例/%
电子期刊库	Elsevier ScienceDirect	1 605	2 369	67.75
	Journal Storage	1 198	3 986	30.06
	Wiley Online Library	1 194	1 446	82.57
	Springer 电子期刊	1 186	1 953	60.73
	Taylor & Francis 期刊数据库	1 077	2 356	45.71
	SAGE 现刊数据库	325	587	55.37
	Project MUSE E-journal	312	639	48.83
全文数据库	Academic Search Premier	6 112	16 387	37.30
	Academic Search Complete	6 082	15 810	38.47
	H. W. Wilson 数据库	4 008	20 322	19.72

表 13　DRAA 2018 中 JCR 期刊数量倒数 10 名的源数据库统计

数据库类别	数据库名	JCR 期刊数量/种	期刊总数/种	JCR 期刊比例/%
电子期刊库	PNAS 数据库	1	1	100.00
	SAE 国际汽车工程师学会数据库	1	16	6.25
	Thieme E-journals	1	37	2.70
	Science Online	5	7	71.43
	AIAA 美国航空航天学会数据库	7	17	41.18
	SPIE Digital Library	7	10	70.00
	APS 全文电子期刊数据库	8	22	36.36
全文数据库	World Bank	0	3	0.00
	TWS 台湾学术期刊在线数据库	34	1 738	1.96
	EBSCO 食品科学全文数据库	422	867	48.67

表 14　DRAA 2018 中 JCR 期刊比例正数前 10 名的源数据库统计

数据库类别	数据库名	JCR 期刊数量/种	期刊总数/种	JCR 期刊比例/%
电子期刊库	PNAS 数据库	1	1	100.00
	SIAM E-journals Package	16	17	94.12
	Nature 研究月刊/评论月刊	45	49	91.84
	IWA	11	12	91.67
	The JAMA Network	11	12	91.67
	AIP 全文电子期刊及会议录数据库	22	24	91.67
	Cell Press	16	18	88.89
全文数据库	ProQuest Biological Science Collection	2 096	3 114	67.31
	PHMC/PML	1 834	3 480	52.70
	ProQuest Agricultural Science Collection	1 923	3 818	50.37

表 15　DRAA 2018 中 JCR 期刊比例倒数 10 名的源数据库统计

数据库类别	数据库名	JCR 期刊数量/种	期刊总数/种	JCR 期刊比例/%
电子期刊库	Thieme E-journals	1	37	2.70
	SAE 国际汽车工程师学会数据库	1	16	6.25
	Bentham Science	41	225	18.22
	Emerald 全文期刊库	57	266	21.43
	Begell Digital Library	11	42	26.19
	Journal Storage	1 198	3 986	30.06
	APS 全文电子期刊数据库	8	22	36.36
全文数据库	World Bank	0	3	0.00
	TWS 台湾学术期刊在线数据库	34	1 738	1.96
	ABI/INFORM Collection	648	7 099	9.13

通过对以上表格数据进行对比可以发现,DRAA 2018 的 43 个电子期刊库中,没有一个电子期刊库能够在 JCR 期刊数量和比例中同时占据前 7 名,而 SAE 国际汽车工程师学会数据库、Thieme E-journals、以及 APS 全文电子期刊数据库同时占据了 JCR 期刊数量和比例后 7 名的位置。同样,DRAA 2018 的 14 个全文数据库中,没有一个全文数据库能够在 JCR 期刊数量和比例中同时占据前 3 名,而 World Bank 和 TWS 台湾学术期刊在线数据库同时占据了 JCR 期刊数量和比例后 3 名的位置。

4.4　DRAA 2018 中各库的独有刊数量和比例

利用以上的映射结果,同样可以很方便地量化 DRAA 2018 中各源数据库的独有刊数量和比例。其主要算法是先求出 DRAA 2018 中某源数据库的 jid 集合与 DRAA 2018 其余源数据库的 jid 集合的差集,然后对差集求模,得到该源数据库库的独有刊数量,再把源

数据库的独有刊数量除以源数据库的期刊总数，即得源库的独有刊比例。限于篇幅，表 16 和表 17 分别列出独有刊数量正数前 10 名和倒数前 10 名 DRAA 2018 源数据库，表 18 和表 19 分别列出了独有刊比例正数前 10 名和倒数前 10 名 DRAA 2018 源数据库。在计算独有刊指标时，参与比较的数据库没有区分电子期刊库和全文数据库，但在展示排名时，还是按 7:3 的比例分别列出了电子期刊库和全文数据库各自的结果。

表 16　DRAA 2018 中独有刊数量正数前 10 名的源数据库统计

数据库类别	数据库名	独有刊数量/种	期刊总数/种	独有刊比例/%
电子期刊库	Springer 电子期刊	740	1 953	37.89
	Elsevier ScienceDirect	735	2 369	31.03
	Journal Storage	592	3 986	14.85
	Taylor & Francis 期刊数据库	372	2 356	15.79
	SAGE 现刊数据库	206	587	35.09
	LWW 医学全文期刊数据库	185	289	64.01
	Project MUSE E-journal	129	639	20.19
全文数据库	H. W. Wilson 数据库	5 894	20 322	29.00
	ABI/INFORM Collection	2 462	7 099	34.68
	ProQuest Research Library	858	6 111	14.04

表 17　DRAA 2018 中独有刊数量倒数 10 名的源数据库统计

数据库类别	数据库名	独有刊数量/种	期刊总数/种	独有刊比例/%
电子期刊库	APS 全文电子期刊数据库	0	22	0.00
	IWA	0	12	0.00
	PNAS 数据库	0	1	0.00
	The JAMA Network	0	12	0.00
	British Medical Journals	1	31	3.23
	Science Online	1	7	14.29
	AGU 美国地球物理学会数据库	2	18	11.11
全文数据库	Business Search Complete	0	6 208	0.00
	Business Source Premier	0	6 208	0.00
	World Bank	3	3	100.00

从以上 4 个表格可以看到，DRAA 2018 的 43 个电子期刊库中，没有一个电子期刊库能够在独有刊数量和比例中同时占据前 7 名，而 APS 全文电子期刊数据库、IWA、PNAS 数据库、The JAMA Network、以及 British Medical Journals 同时占据了独有刊数量和比例后 7 名的位置。在 DRAA 2018 的 14 个全文数据库中，H. W. Wilson 数据库与 ABI/INFORM Collection 在独有刊数量和比例中同时占据了前 3 名，Business Search Complete 与 Business Source Premier 同时占据了独有刊数量和比例后 3 名的位置。

表 18　DRAA 2018 中独有刊比例正数前 10 名的源数据库统计

数据库类别	数据库名	独有刊数量/种	期刊总数/种	独有刊比例/%
电子期刊库	Thieme E – journals	36	37	97.30
	Begell Digital Library	38	42	90.48
	RSC 英国皇家化学学会期刊及数据库	39	55	70.91
	LWW 医学全文期刊数据库	185	289	64.01
	American Society for Microbiology	7	12	58.33
	Bentham Science	125	225	55.56
	Independent Scholarly Publishers Group	17	31	54.84
全文数据库	World Bank	3	3	100.00
	ABI/INFORM Collection	2 462	7 099	34.68
	H. W. Wilson 数据库	5 894	20 322	29.00

表 19　DRAA 2018 中独有刊比例倒数 10 名的源数据库统计

数据库类别	数据库名	独有刊数量/种	期刊总数/种	独有刊比例/%
电子期刊库	APS 全文电子期刊数据库	0	22	0.00%
	IWA	0	12	0.00%
	PNAS 数据库	0	1	0.00%
	The JAMA Network	0	12	0.00%
	British Medical Journals	1	31	3.23%
	Wiley Online Library	75	1 446	5.19%
	AIP 全文电子期刊及会议录数据库	2	24	8.33%
全文数据库	Business Search Complete	0	6 208	0.00%
	Business Source Premier	0	6 208	0.00%
	Academic Search Complete	22	15 810	0.14%

5　算法分析

采用统一期刊空间中的多库比较算法，主要有以下几个优点：

（1）一次映射，多次比较。一个数据库一旦映射到统一的期刊空间后，就可以与同样映射到此期刊空间的其他源库或参考库，进行任意的组合比较，无须根据比较库的不同而重新调整比对策略。

（2）准确率高。只利用 ISSN 和 EISSN 对期刊进行比对的算法，忽略了 ISSN 和 EISSN 都为空值的期刊；只利用规范化刊名对期刊比对的算法，很容易把不同刊的刊名规范成一样的刊名，容易造成统计误差的扩大。利用 ISSN、EISSN 和刊名字段，但在刊名比对时没有对刊名进行预处理或者只进行了简单的预处理，也容易扩大后续的统计误差。

庄纪林，陈凌，姚晓霞.外文期刊数据库内容质量指标的量化研究[J].图书情报工作,2021,65(2):54-64.

(3)有利于对非刊记录与错误记录的排除。在数据库商提供的期刊列表中,发现错误是毫不奇怪的事情,而把这些期刊统一映射到同一个期刊空间中,有助于把一些非刊记录和错误记录剔除出去。虽然此期刊空间的期刊记录也不能保证是完备的,但是由于所有参与比较的数据库都统一映射到该空间,所以避免了库与库之间单独比较时采用不同的标准。

当然,采用统一期刊空间中的多库比较算法,相关指标量化的正确性依赖于统一期刊空间的质量。如果统一期刊空间中的数据正确率不高、数据不全、或者一致性较差,也会给指标的量化带来一定的误差。

6 结语

独有刊由于比较范围的不同,可分为绝对独有刊和相对独有刊。有些期刊由出版社独家授权给数据库提供商发行,则这些期刊对数据库商来说是绝对独有刊。但是对于图书馆来说,追求的不仅仅是绝对独有刊,相对独有刊也是很重要的指标。相对独有刊的比较范围限定在图书馆的电子资源采购目录或某一特定的范围之内。数据库商能给图书馆提供的是绝对独有刊的数量,但是相对独有刊的数量就需要图书馆自己量化了。采用本研究的成果,运用基于同一期刊空间的多库比较算法,则很容易计算相对独有刊,同时也很方便在增加或去除一个或多个数据库之后,考察独有刊分布的调整。笔者对 DRAA 2018 的 57 个源数据库计算独有刊指标,实际上也是相对独有刊指标的计算。其值通常大于绝对独有刊,所选定的数据库范围越大,则相对独有刊指标越接近于绝对独有刊指标。

在此基础上,同样可以衍生更多内容质量方面的二级指标,如全文刊、JCR 全文刊、JCR 现刊、SNIP 全文刊、SNIP 现刊、OA 全文刊、OA 现刊、独有全文刊、独有现刊等。只要获取到合适的数据,利用本研究的算法,可以很方便地计算出这些指标。

外刊数据库采购是图书馆资源经费使用的重要组成部分,而内容质量是外刊数据库评价指标中的重中之重,精确量化其中的二级指标对采购决策意义重大。本研究针对外文期刊数据库内容质量方面的指标量化算法,提出了在统一期刊空间中进行期刊比对的方法,并对 JCR 期刊数量/比例与独有刊数量/比例做了实证研究,以期为图书馆的资源建设和推广利用提供有力的支撑,使得馆藏资源的建设更加合理、经济。

参考文献:

[1] 高校图书馆数字资源采购联盟(DRAA). 资源百科[EB/OL]. [2020-07-30] http://www.libconsortia.edu.cn/Ency/index-Ency.action.

[2] BLIXRUD J C. Measures for electronic use: The ARL E-Metrics project[M]// Statistics in practice - measuring & managing. Loughborough: IFLA satellite conference, 2002: 73-84.

[3] COUNTER. Counter code of practice: release 5[EB/OL]. [2020-11-20]. https://www.projectcounter.org/wp-content/uploads/2017/11/Release5_20171107.pdf.

[4] BROPHY P. Electronic library performance indicators: the EQUINOX project[J]. Serials, 2006, 14(1): 5-9.

[5] International organization for standardization. Information and documentation - International library statistics: ISO 2789: 2013 [S]. Switzerland: ISO, 2013: 4-16.

[6] JOHN CARLO BERTOT, CHARLES R. MCCLURE, JOE RYAN. Developing national network statistics and performance measures for US public libraries: issues, and recommendations [J]. Performance measurement and metrics, 2000, 1(1): 15-41.

[7] NISO. ANSI/NISO Z39.7-2013 information services and use: metrics & statistics for libraries and information providers data dictionary[EB/OL]. [2020-07-30]. https://www.niso.org/publications/z397-2013.

[8] SUTTON S. A model for electronic resources value assessment[J]. Serials librarian, 2013, 64(1-4): 245-253.

[9] KINMAN V. E-Metrics and library assessment in action[J]. Journal of electronic resources librarianship, 2009, 21(1): 15-36.

[10] TIMMS G. Gathering, evaluating, and communicating statistical usage information for electronic resources[M]// Managing electronic resources: a LITA guide. Chicago :ALA TechSource, 2012: 89-120.

[11] WALKER M. E-Resource statistics: what to do when you have no money[J]. Journal of electronic resources librarianship, 2009, 21(3/4): 237-250.

[12] 肖珑,李浩凌,徐成. CALIS 数字资源评估指标体系及其应用指南[J]. 大学图书馆学报,2008(3):2-8,17.

[13] 肖珑,张宇红. 电子资源评价指标体系的建立初探[J]. 大学图书馆学报,2002(3):35-42,91.

[14] 姚晓霞,陈凌. 图书馆业绩评估国际标准及其应用——ISO11620 与 EQUINOX[J]. 图书情报工作,2003(2):98-101,71.

[15] 向英明,谭艺曼,林欢. 电子资源综合评价指标体系及其数学模型的研究[J]. 图书馆杂志,2004(1):26-29.

[16] 徐革. 电子资源评价之重要影响因子的调查研究[J]. 大学图书馆学报,2006(3):77-81,96.

[17] 刘军,莫利江,吴朗,等. 电子资源综合评价指标体系的构建探讨[J]. 情报杂志,2010,29(S1):135-137.

[18] 唐琼. 采用因子分析法构建数字资源选择标准体系[J]. 大学图书馆学报,2011,29(2):56-63.

[19] 和媛媛,巩在武. 基于模糊多属性决策的高校图书馆电子资源服务绩效评价研究[J]. 情报科学,2011,29(3):412-415,

420.
[20] 刘文梅, 唐淑娟. 我国“211 工程”高校数字图书馆绩效评估研究——已有研究提出的指标在我国的适用性[J]. 图书情报知识, 2006 (5) :13 -18.
[21] 索传军. 数字馆藏服务绩效评估指标体系及其构建原则[J]. 图书情报知识, 2006 (5) :13 -18.
[22] 李新霞. 中外数字图书馆绩效评估的比较研究[J]. 图书馆学研究,2013(7):34 -41.
[23] 张轶华,李芳,宋海艳,等. 基于实证比较的外文期刊数据库评价模型构建与应用[J]. 图书馆杂志,2019,38(8):65 -73,16.
[24] ProQuest. ULRICHSWEBTM global serials directory[EB/OL]. [2020 -07 -30]. https://ulrichsweb.serialssolutions.com/.

作者贡献说明:
庄纪林:处理与分析数据,撰写和修改论文;
陈凌:提出问题,设计论文研究思路和结构;
姚晓霞:审核与修改论文。

Quantitative Research on Content Quality Indicators of Foreign Journal Database

Zhuang Jilin Chen Ling Yao Xiaoxia
Peking University Library, Beijing 100871

Abstract: [Purpose/significance] Aiming at the quantification of double-base and multi-base indicators of content quality of foreign journal database, a multi-base comparison algorithm based on unified journal space is proposed, which realizes the journal identification and comparison between arbitrary multi-bases, and improves the quantification accuracy of these indicators. [Method/process] On the basis of analyzing the content and complexity of the content quality indicators of foreign journal databases, this paper designed a multi-database comparison algorithm based on unified journal space, and used this algorithm to calculate the number of JCR journals and the number of unique journals of 57 databases in DRAA in 2018, as a verification of the validity and accuracy of the algorithm. [Result/conclusion] As an extension study of various electronic resource evaluation indicator systems, this study can help libraries optimize their collection construction and make the purchase of foreign journal databases more economical and reasonable.

Keywords: foreign journals database evaluation content quality indicator quantification unified journal space